The Report Committee for Lin Qiu
Certifies that this is the approved version of the following report:

Methods of Genotype Imputation for Genome-wide Association Studies

APPROVED BY
SUPERVISING COMMITTEE:

Miachel Daniels, Supervisor

Lizhen Lin

Methods of Genotype Imputation for Genome-wide Association Studies

by

Lin Qiu, B.S.;M.Ag.

Report

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN STATISTICS

The University of Texas at Austin

August 2016

Abstract

Methods of Genotype Imputation for Genome-wide Association Studies

Lin Qiu, M.S.Stat.

The University of Texas at Austin, 2016

Supervisor: Miachel Daniels

In genetic epidemiological studies, missing data problems arise when genotypes of particular markers are unavailable for reasons of data quality, cost efficiency or technical design. Genotype imputation is a well-established statistical technique for estimating unobserved genotypes in association studies. Imputation methods are implemented by copying haplotype segments from a densely genotyped reference panel into individuals typed at a subset of the reference variants. By this way, genotypes can be estimated and tested for association at variants that were not assayed in a study. This report first summarizes the missing data mechanisms. Then an overview of the different methods that have been proposed for genotype imputation is provided and some thoughts for future directions are given.

Table of Contents

Abstract iv

List of Tables vi

List of Figures vii

Chapter 1. Missing Data 1

1.1 Introduction 1

1.2 Sources of Missing Data 1

1.3 Missing Mechanisms 2

Chapter 2. Missing Data in Genetic Epidemiology 4

2.1 Introduction 4

2.2 Imputation Methods 5

2.2.1 IMPUTE 5

2.2.2 fastPHASE 10

2.2.3 MaCH 12

2.2.4 BEAGLE 13

Chapter 3. Comparsion Between Imputation Methods 16

Chapter 4. Discussion and Future Directions 20

Bibliography 23

List of Tables

Table 1: Reference properties 17

Table 2: Properties of the study sample 17

Table 3: Properties of program options and features 18

Table 4: Computational performance 19

List of Figures

Figure 1: Genotype imputation with IMPUTE 8

Figure 2: Genotype imputation with fastPHASE.....12

Figure 3: Genotype imputation with BEAGLE 15

Chapter 1

Missing Data

1.1 Introduction

In statistics, missing data, occur when no data value is stored for the variable in an observation. Missing data are a common occurrence and have challenged researchers since the beginnings of field research, particularly for longitudinal research, which involves multiple waves of measurement on the same individuals. The procedures researchers use were mainly developed in the twentieth century which designed for complete data.

Missing data can adversely impact the validity of statistical inferences.

1.2 Sources of Missing Data

There are three main sources of missing data in survey research: noncoverage, total nonresponse, and item nonresponse (Groves et al., 2004). Noncoverage happens when some population have no chance of being selected in the sample. Missing data comes from nonresponse when a respondent refuses to respond to any item on the survey. Item nonresponse occurs when only partial of the items on the survey is being completed.

Among these three sources of missing data, noncoverage and total nonresponse can be addressed by using appropriate sampling weights that are designed to make the sample accurately represent the target population (Cheema, 2014). However, missing values due to item nonresponse cannot be fixed by using weights. Listwise deletion of cases with item nonresponse will result in loss of some of the valuable information and potentially lead to bias. Missing value imputations account for the uncertainty in the missing value.

A main concern with missing data is whether the sample with complete data is still representative of the target population (Roth, 1994). Let's take an example of a sample of 100 high school students of whom 5 are from school A. If data on A school are missing, at 5%, the overall missingness rate is small. In addition sample from the other schools was not representative of its target population because if does not contain any of A school. If school A is the 'same' as the other schools, there will be no problem. However, we cannot check this since we don't observe data from school A.

1.3 Missing Mechanisms

Missing data mechanisms are typically categorized into three groups (Rubin, 1976): missing completely at random (MCAR), missing at random (MAR), and missing not at random (MNAR).

If subjects who have missing data are a random subset of the complete sample of subjects, missing data are called completely at random (MCAR). When missing data are MCAR, the set of subjects with no missing data is a random sample from the target

population. Using only complete data will give unbiased results. However, such an analysis is less efficient.

MAR occurs when when missingness does not depend on the observed data.

Data are MNAR when the probability of missing data on a variable depends on the value of that variable.

Chapter 2

Missing Data in Genetic Epidemiology

2.1 Introduction

In genetic epidemiological studies, missing data problems arise when genotypes of particular markers are unavailable for analysis for reasons of data quality, cost efficiency or technical design. Imputation methods work by using haplotype patterns in a reference panel to predict unobserved genotypes in an experimental sample. To make inference on an observed genotypes using unobserved genotypes we use information from a set of comprehensively typed individuals (like HapMap or the 1000 Genomes Project). The most common genotype imputation methods including IMPUTE, fastPHASE, MaCH and BEAGLE (Marchini & Howie, 2010).

Human genetic epidemiology aims to identify genetic variants related to specific phenotypes, usually diseases or disease-associated traits. Genotyping studies have been always confined to selected marker sets, mainly single nucleotide polymorphisms (SNPs) that are formatted on commercially available microarrays. Genome-wide association studies (GWAS) based on linkage disequilibrium (LD), only use a small number of SNPs that characterize 80% of the genetic variation in a given population (Krawczak, 2015). Depends on how they are chosen, SNPs might be proxies of truly causative mutations. Unfortunately, commonly used microarray-based SNPs genotyping result in errors.

Genotype imputation methods can help scientists address the missing data problem (Marchini & Howie, 2010).

2.2 Imputation Methods

Various genotype imputation methods have been proposed for epidemiological studies. The main difference between those genotype imputation methods is how the conditional genotype distribution of an SNP is defined and used. Assuming we have data at *L* diallelic autosomal SNPs and that the two alleles at each SNP have been coded 0 and 1. H denotes a set of N haplotypes at these L SNPs while G is the set of genotype data at the L SNPs. The purpose of genotype imputation is to predict the genotypes of those SNPs that have not been genotyped in the study sample. All of those imputation methods discussed here assume the missingness is MAR and are HMM based imputation methods (including IMPUTE v1). They infer the missing genotypes for each study individual independently, conditional on the reference panel. By doing so, the reference panel is used to analytically integrate over the phase uncertainty in each individual's multilocus genotype.

2.2.1 IMPUTE

Marchini *et al*. (2007) devised the first version of IMPUTE method (Figure 1). IMPUTEv1 is based on an extension of the hidden Markov models (HMMs) for simulating coalescent trees (Fearnhead & Donnelly, 2001) and for modeling of linkage

disequilibrium and estimating of recombination rates (Li & Stephens, 2003). The method is based on an HMM of each individual's genotype vector G_i conditional upon a set of N known haplotypes H. A Hidden Markov Model (HMM) has the form

$$\Pr(G_i \mid H) = \sum_{Z_i^{(1)} Z_i^{(2)}} \Pr(G_i \mid Z_i^{(1)}, Z_i^{(2)}, H) \Pr(Z_i^{(1)}, Z_i^{(2)} \mid H), \quad (1)$$

where $Z_i^{(1)} = \left\{ Z_{i1}^{(1)}, ..., Z_{iL}^{(1)} \right\}$ and $Z_i^{(2)} = \left\{ Z_{i1}^{(2)}, ..., Z_{iL}^{(2)} \right\}$ are two sequences of hidden states at the L sites and $Z_{il}^{(j)} \in \left\{ 1, ..., N \right\}$. The Z_i can be thought of as the pair of haplotypes from the reference panel H that are being copied to form the genotype vector G_i. The term $\Pr(Z_i^{(1)}, Z_i^{(2)} \mid H)$ defines the prior probability on how the pair of copied haplotypes changes along the sequence and is defined by a Markov chain in which the switching rates depends on an estimate of the fine-scale recombination map across the genome. The initial state of the Markov chain is uniform on the N^2 states

$$\Pr(Z_{i1}^{(1)}, Z_{i1}^{(2)} \mid H) = \frac{1}{N^2}, \quad (2)$$

We can compute the transition probabilities of the chain from site l to $l+1$ are given by

$$\Pr(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \to \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\} \mid H) = \begin{cases} \left(e^{-\frac{\rho_l}{N}} + \frac{1 - e^{\frac{\rho_l}{N}}}{N} \right)^2 & Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ \left(e^{-\frac{\rho_l}{N}} + \frac{1 - e^{\frac{\rho_l}{N}}}{N} \right) \left(\frac{1 - e^{\frac{\rho_l}{N}}}{N} \right) & Z_{il}^{(1)} = Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \\ & Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} = Z_{i(l+1)}^{(2)} \\ \left(\frac{1 - e^{\frac{\rho_l}{N}}}{N} \right)^2 & Z_{il}^{(1)} \neq Z_{i(l+1)}^{(1)}, Z_{il}^{(2)} \neq Z_{i(l+1)}^{(2)} \end{cases} \tag{3}$$

$\rho_l = 4N_e r_l$ and r_l is the per generation genetic distance between sites l and $l+1$. The prior distribution could be written as

$$\Pr(Z_i^{(1)}, Z_i^{(2)} \mid H) = \Pr(Z_{i1}^{(1)}, Z_{i1}^{(2)} \mid H) \prod_{l=1}^{L-1} \Pr(\{Z_{il}^{(1)}, Z_{il}^{(2)}\} \to \{Z_{i(l+1)}^{(1)}, Z_{i(l+1)}^{(2)}\} \mid H), \tag{4}$$

The term $\Pr(G_i \mid Z_i^{(1)}, Z_i^{(2)}, H)$ defines how the observed genotypes will be close to but not exactly the same as the haplotypes being copied.

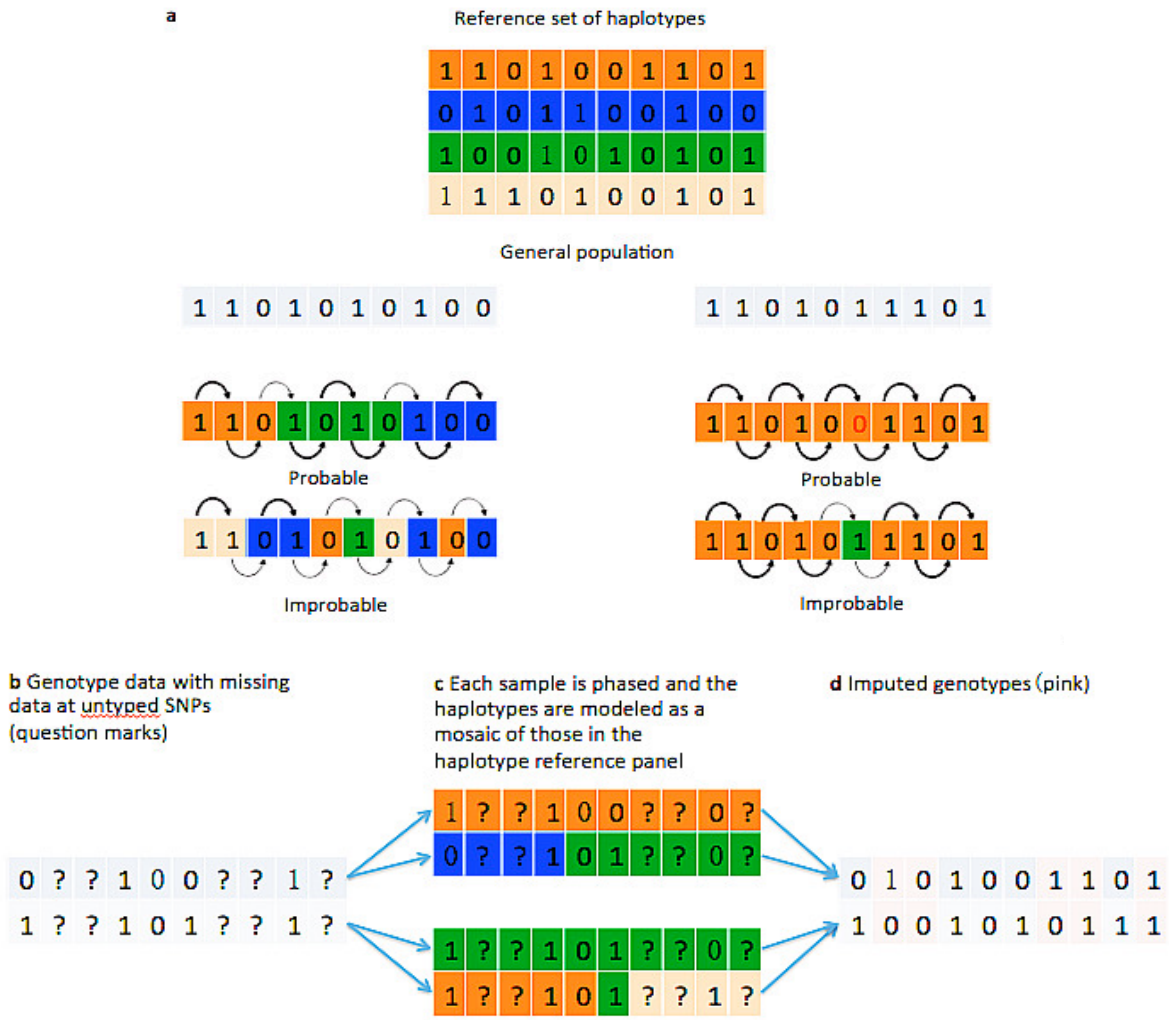

Figure 1 Genotype imputation with IMPUTE based on Krawczak (2015). 0 and 1 indicating the presence or absence of a reference allele, the imputation base (reference set of haplotypes) is phased and contains four different haplotypes. (a) The population haplotype is assumed to be a mosaic of haplotypes from the reference set. Defined by Markov chain model with transition probabilities, the respective haplotype distribution depends both on the population history and the local recombination map. Bold arrows indicate high transition probabilities; while think arrows represent lower transition probabilities. Red means mutations. (b) Genotype data with missing data at untyped SNPs(question marks). (c) Each sample is phased and the haplotypes are modeled as a mosaic of those in the haplotype reference panel. (d) Missing genotypes in the study sample are imputed using the matching haplotypes in the reference set. In general, the probability of an un-phased genotype with missing data is evaluated by considering all possible pairs of mosaic haplotypes that would be compatible with the observed data. And the most probable pair determines the most probable genotypes at un-typed SNPs.

IMPUTEv2 developed by Li & Stephens (2003), a more flexible approach than IMPUTEv1 was used. Rather than perform a separate, analytical imputation step for each individual, IMPUTEv2 imputes all individuals together in an iterative framework. SNPs are divided into two disjoint sets based on whether an SNP was genotyped in both the study sample and reference panel (set T) or just was genotyped in the reference panel (set U). The haplotypes at SNPs in the study sample (set T) are estimated first and then alleles at SNPs in U are imputed conditional on the current estimated haplotypes. Specifically, it runs a Markov chain Monte Carlo (MCMC) algorithm that alternates between two basic steps: 1) Phase all observed genotypes and impute any sporadically missing genotypes at SNPs in the study sample (SNPs in the set T). Pool phase information across all individuals that are genotyped at a given SNP; 2) For each haplotype inferred in the previous step, use the reference panel to impute missing alleles at untyped SNPs (Howie & Marchini, 2009). A probability distribution for the haplotypes is defined similar to IMPUTEv1, and the two haplotypes have the highest posterior probability are selected for the study sample. IMPUTEv2 is much faster compared to IMPUTEv1 since the imputation step is haploid imputation, reducing computation time from O (N^2) in IMPUTEv1 to O (N) in IMPUTEv2(N denotes the number of possible haplotypes) (Howie, Donnelly & Marchini, 2009 ; Marchini & Howie, 2010).

2.2.2 fastPHASE

This method is based on SNP haplotypes tend to cluster into groups with similar haplotypes (Sheet & Stephens, 2006), and has been implemented in an association-testing program called BIMBAM (Servin & Stephens, 2007). This model specifies a set of K unobserved states to represent common haplotypes and each cluster (kth) is assigned a weight (a_{kl}) proportional to the fraction of haplotypes contained in each cluster at site *l*.

$$\sum_{k} a_{kl} = 1 , \tag{5}$$

θ_{kl} is the frequency of allele 1 at each site in each cluster. fastPHASE models population haplotypes as an HMM with transitions between states.

$$P(G_i \mid \alpha, \theta, r) = \sum_{z} P(G_i \mid Z_i, \theta) P(Z_i \mid \alpha, r) , \tag{6}$$

$P(G_i \mid Z, \theta)$ defines how likely the observed genotypes, and $P(Z \mid \alpha, r)$ models patterns of switching between states which represent clusters not the reference haplotypes. And here, a likelihood can be written as

$$L(G, H \mid \alpha, \theta, r) = \prod P(G_i \mid \alpha, \theta, r) \prod P(H_i \mid \alpha, \theta, r) , \tag{7}$$

An EM algorithm is used to fit the model, and a forward-backward algorithm is used to impute the missing genotypes conditional on the parameter estimation. Researchers found that by averaging a set of estimates resulted much better results than just a single estimate. fastPHASE uses smaller sets of states which should reasonably be computationally faster, however, this advantage is partly outweighed by working with abstract clusters requires many parameters. And also, in fastPHASE more unknown parameters are involved in the likelihood calculations than in IMPUTE, plus the effect of cluster weights, fastPHASE has greater standard errors on parameter estimaties. Fixing some of the parameters at values obtained from a phased imputation base could be a possible solution (Marchini & Howie, 2010).

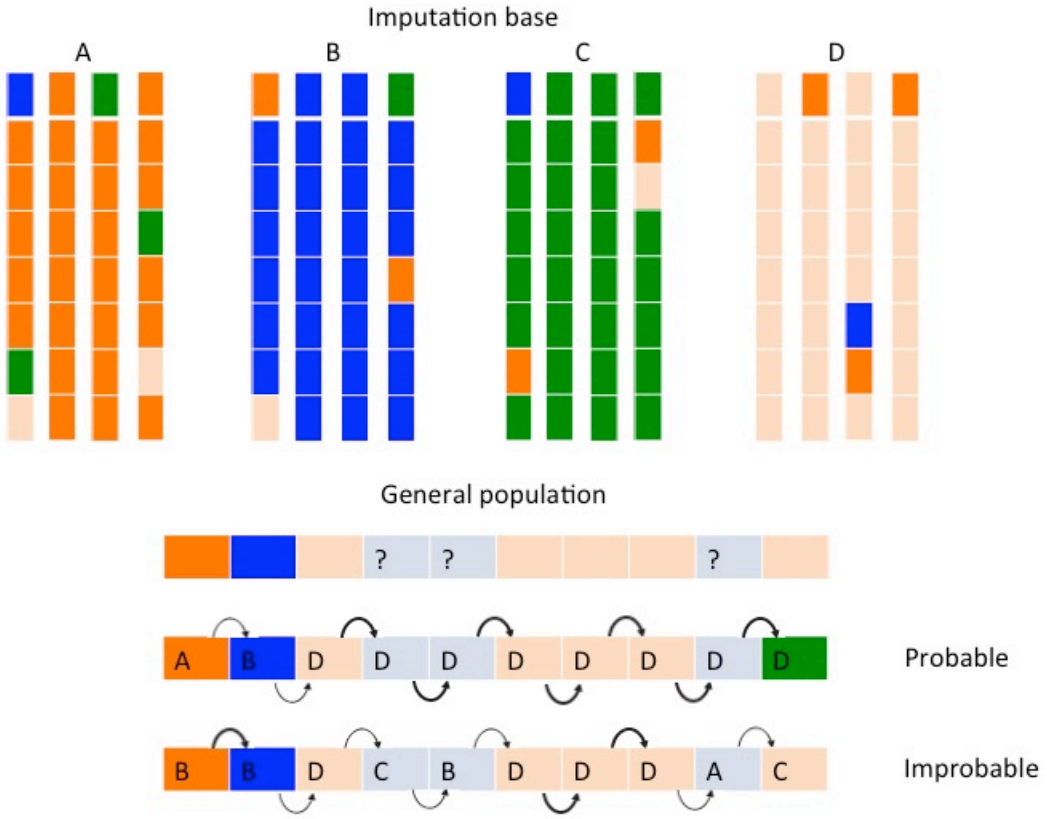

Figure 2 Genotype imputation with fastPHASE based on Krawczak (2015). Haplotypes are assumed to cluster. A,B,C and D define different sets of the Markov chain generating the haplotype distribution. The alleles are color-coded, and different colors correspond to identical alleles at a given position.

2.2.3 MaCH

MaCH employs an HMM model which is very similar to IMPUTE, but, in contrast to IMPUTE and fastPHASE, MaCH does not require a separate imputation base, by constantly updates the phase of each individual's genotype data conditional on the current haplotype estimates of all the other samples. This model can be written as

$$P(G_i \mid D_{-i},\theta,\eta) = \sum P(G_i \mid Z,\eta)P(Z \mid D_{-i},\theta)\ , \tag{8}$$

D_{-i} is the set of estimated haplotypes except I, Z is the hidden states of HMM, w η determines how similar G_i is to the copied haplotypes, and θ controls transitions between the hidden states(Marchini & Howie, 2010). The population haplotype distribution is determined iteratively and parameters η and θ are also updated during each iteration.

Based on a reference panel of haplotypes, H, imputation of unobserved genotypes is accommodated by adding H to the estimated haplotypes D_{-i}. Then through the haplotypes sampled iteratively the marginal distribution of the unobserved genotypes can be estimated.

2.2.4 BEAGLE

The BEAGLE method (Browning, 2006) forms an HMM by locally clustering the haplotypes at each marker position along a chromosome. The clusters are defined at the allele not the haplotype level. Another difference with fastPHASE is that cluster number is region-dependent not fixed. The model has no parameters that need to be estimated and it is applied to a given set of haplotypes in two steps. The local level of allelic association between proximal markers determines the transition probabilities rather than global haplotype frequencies.

There is no need to model the historical recombination process since Markov chain has no 'long-term memory', and no mutation needs to be allowed because the sole states of the Markov chain in BEAGLE are truly observed alleles. All of these features make BEAGLE computationally faster in genotype imputation compared to other methods mentioned above. However, the drawback is that BEAGLE has to be based on full study sample which means it cannot be split up like IMPUTE2. There will automatically be assigned a probability of zero if any haplotype containing an unaccounted allele combination at neighbouring markers.

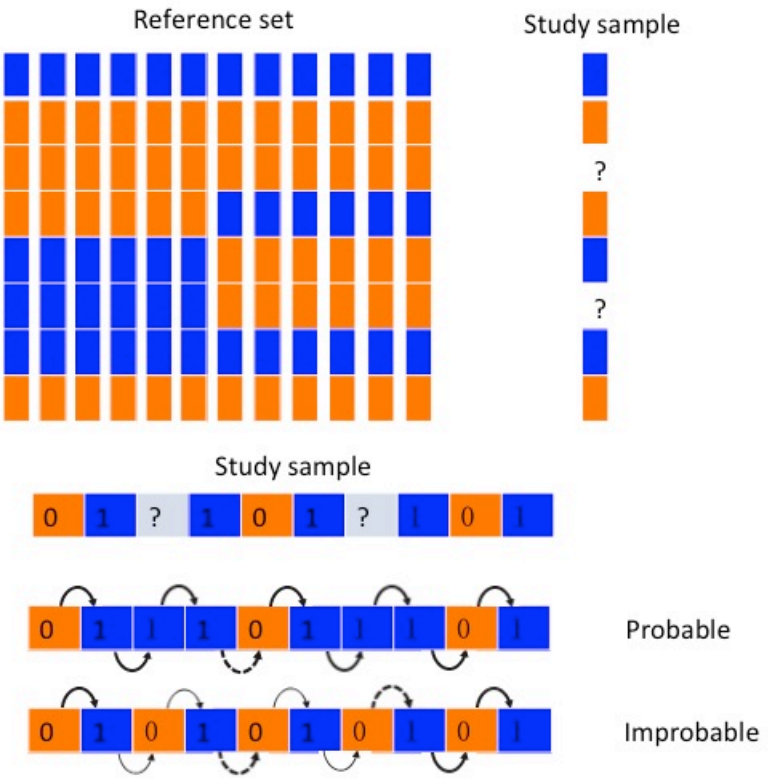

Figure 3 Genotype imputation with BEAGLE. Different from fastPHASE based on Krawczak (2015), BEAGLE method defines clusters at the allele rather than the haplotype level, for SNPs, the cluster number equals two at each position (blue or orange). Both the local level of inter-marker allelic association in the imputation base and study sample determines the haplotype distribution probabilities. Broken arrows indicate the Intermediate transition probabilities.

Chapter 3

Comparison Between Imputation Methods

Researchers have carefully studied the performance of these commonly used genotype imputation frameworks (Marchini & Howie, 2009 ; Nothnagel et al., 2009 ; Wang et al.,2012 ; Liu et al.,2014).

Marchini & Howie (2010) summarized the properties of each of the most popular imputation methods according to the reference panels the methods can handle (Table 1), properties of the study samples (Table 2), program options (Table 3), computational performance and error rates (Table 4). They found that IMPUTE v2 is the most accurate approach among IMPUTEv1, MACH, fastPHASE, and BEAGLE based on their simulation scenario and under a MAR assumption, but all the methods produce similar performance. Moreover, to further examine how methods perform on a large reference panel of haplotypes, like 1000 Genomes Project, the author used a reference panel of 1,000 haplotypes consisting 500 and 1,000 individuals and applied HAPGEN simulation based on pilot CEU haplotypes from the 1000 Genomes Project in a 5 Mb region on chromosome 10. The results showed IMPUTEv2 is faster compared to BEAGLE and fastPHASE.

Properties	Imputation method				
	IMPUTEv1	IMPUTE v2.2	MACH V1.0.16	fastPHASE v1.4.0	BEAGLE v3.2
Reference panels					
Can use a haplotype reference panel?	Yes	Yes	Yes	Yes	Yes
Can use a genotyped reference panel?	No	Yes	Yes	Yes	Yes
Can two haplotype or genotype reference panels be used in the same run?	No	Yes	No	No	No
Reference panels available in correct format	HapMap2 HapMap3 1KGP pilot data	HapMap2 HapMap3 1KGP pilot data 1000 Genome projects	HapMap2 HapMap3 1KGP pilot data 1000 Genome projects	Hap2	1000Genome projects

Table 1 Reference properties.

Properties	Imputation method				
	IMPUTEv1	IMPUTE v2.2	MACH v1.0.16	fastPHASE v1.4.0	BEAGLE v3.2
Study samples					
Can take genotypes specified with uncertainty?	No	Yes	No	No	Yes
Can accommodate trios and related samples?	No	No	No	No	Trios and duos
Can impute into a study sample of autosomal haplotypes?	Yes	Yes	No	No	Yes
Can impute on the X chromosome?	Yes	Yes	No	No	Yes

Table 2 Properties of the study sample.

Wang *et al*. (2012) proposed a new nearest neighbor method (NN) and a weighted variant (WNN) which both follow the coalescent theory that the target individual has a genotype sequence similar to one from the population. Besides these two methods, they also implemented fastPHASE, Npute (Roberts et al., 2007) and several machine learning imputation methods including support vector machine (SVM), a local neural network (NeuralNet), and a local first order Markov chain (MC). The results showed NN and WNN were among the most efficient methods, and performed much better than other methods except fastPHASE in missing SNP genotype imputation.

Properties	Imputation method				
	IMPUTEv1	IMPUTE v2.2	MACH v1.0.16	fastPHASE v1.4.0	BEAGLE v3.2
Program options and features					
Does phasing as well as imputation?	No	Yes	Yes	Yes	Yes
Can impute sporadic missing genotypes?	No	Yes	Yes	Yes	Yes
Has internal performance assessment?	Yes	Yes	Yes	No	No
Can impute only in a specified interval?	Yes	Yes	No	No	No
Can handle strand alignment between data sets?	Yes	Yes	Yes	No	No
SNP and sample inclusion and exclusion options?	Yes	Yes	No	Yes	Yes
Joint model for imputation and association testing?	No	No	No	No	No
Operating system requirements?	Linux,Solaris, Windows,Mac	Linux,Solaris, Windows,Mac	Linux,Windows, Mac	Linux,Solaris,Windows,Mac	Java executable

Table 3 Properties of program options and features.

Liu et al. (2014) systematically examined the genotype imputation performance based on whole-genome DNA sequencing data from 90 individuals. By using the percentage of variants that were accurately imputed as the criteria to evaluate imputation performance, firstly, they found that minimac and IMPUTE2 have better imputation performance compared to BEAGLE and multi-population reference panel showed better performance than just using a reference panel from the same population. Second, investigators usually rely on imputation quality measure to remove poorly imputed variants from further analysis.

Properties	**Imputation method**				
	IMPUTEv1	IMPUTE v2.2	MACH V1.0.16	fastPHASE v1.4.0	BEAGLE v3.2
Computational performance					
Assessment1★	43m(100Mb)	75m(180 Mb)	105m(80Mb)	855m(16Mb)	56m(3100Mb)
Assessment2#	---	48m(115m)	---	157m(211m)	104m(234m)
Error ratesζ					
Rows correspond to the Scenario A	5.42%	5.16%	5.46%	5.92%	6.33%
Scenario B(restricted)	---	3.4%(0.86%)	---	5.33%(1.32%)	3.46%(0.93%)
ScenarioB (full)	---	3.4%(0.86%)	---	---	4.01%(1.04%)

Table 4 Computational performance. ★Imputation of 1377 samples on the Affy500k chip from 120 CEU Hapmap2 haplotypes; 7.5 Mb region. Data comes from (Howie, et al., 2009). # Imputation of 500(1000) samples genotyped at 872 SNPs from 1000 haplotypes at 8712 SNPs in a 5 Mb region. Timings based on datasets simulated using HAPGEN and the pilot CEU haplotypes from the 1000 Genomes project in a 5 Mb region on chromosome 10. ζ Error rates from (Howie, et al., 2009), the results for IMPUTEv2 have been updated. Scenario B error rates given are for lllumina SNPs imputed from Affymetrix SNPs. Error rates for Affymetrix SNPs imputed from lllumina SNPs are given in brackets.

Chapter 4

Discussion and Future directions

From the missing data scenarios considered in these studies IMPUTEv2 had both higher accuracy and faster computation. Howie et al. (2011) believes the success of IMPUTE2 is due to its computational strategies and its model of DNA sequence variation. From the perspective of models' algorithm, BEAGLE and fastPHASE combine haplotypes into clusters which speeds up the computation process since it restricts the number of HMM states that need to be involved. BEAGLE and fastPHASE only need to run the calculations on a smaller set of clusters instead of performing HMM calculations on every haplotype in the dataset. By contrast, IMPUTEv2 and MaCH perform HMM for every haplotype in a state; however, using all of the states makes computation intractable, and so IMPUTEv2 restricts the states. The intuition is that the "surrogate family members" identified should include the most informative haplotypes for a particular individual in a particular part of the genome. Both of these state-reduction approaches speed up imputation, and Howie et al. (2011) demonstrated that IMPUTEv2 achieves higher accuracy than BEAGLE for the scenarios they examined, and it is particularly obvious at low-frequency variants in datasets that have higher haplotype diversity. Clustering methods will obscure the differences among those haplotypes which reduces the ability to impute low-frequency variants. As datasets continue to grow more, states

will be needed to add to HMMs for clustering methods to be accurate, IMPUTEv2 will need to be modified to balance accuracy and running time in future practical use.

Another technique for increasing the efficiency of imputation is called "pre-phasing"(Howie et al., 2012) which could maintain accuracy while reducing computational costs.

Genotype imputation is becoming a de rigeur part of genome-wide association studies (GWAS) that have delivered hundreds of bona-fide associations to complex human diseases (Hindorff et al., 2009) and it will continue to be over the next few years. The main factor that will influence which imputation method is used will be those that can handle the increasing availability of next-generation sequencing data with much larger numbers of SNPs. That is also the challenge for imputation methods in using larger, more diverse set of haplotypes.

Certainty needs to be exercised in using imputation methods. First, the most commonly used data from HapMap and 1000 Genomes Project comprise a diverse selection of reference populations. As a consequence, genotype imputation using these resources as reference set (imputation base) may be a valid component of genotype-phenotype association tests since statistical tests evaluate data based on the null hypothesis that cases and controls from the same population. Thus, if the null hypothesis is incorrect, it does not seem sensible to draw inference on the haplotype distribution and to estimate risks from imputed genotypes. Meanwhile, genotype imputation has been suggested to improve the fine mapping of disease association signals. However, p-values calculated for observed genotypes and the imputed genotypes are not directly comparable

(Krawczak, 2015). The p-values for imputed genotyped may be incorrect if the null hypothesis is wrong since imputation base is not representative of the haplotype distribution as discussed above.

Finally, genotype imputation should follow similar rules of good scientific practice like laboratory-based data generation. There have been a lot of efforts made in the past to define criteria for data quality in genetic disease association studies (Anderson et al., 2010). Therefore, it is necessary to develop a similar level of accuracy and reliability evaluation criteria for genotype imputation.

Bibliography

Anderson, C.A., Pettersson, F.H., Clarke,G.M., *et al*. (2010). Data quality control in genetic case-control association studies. *Nature Protocols*, 5:1564-1573.

Browning, S.R. (2006). Multilocus association mapping using variable-length Markov chains. *American Journal of Human Genetics*, 78:273-280.

Cheema, J.R. (2014). A review of missing data handling methods in educational research. *Review of Educational Research*, XX (X):1-22.

Fearnhead P., & Donnelly P. (2001). Estimating recombination rates from population genetic data. *Genetics*, 159: 1299-1318.

Hindorff, L.A. *et al*. (2009). Potential etiologic and functional implications of genome-wide association loci for human diseases and traits. *Proceedings of the National Academy of Sciences,* 106(23): 9362-9367.

Howie, B.N, Donnelly, P., & Marchini, J. (2009). A flexible and accurate method for the next generation of genome-wide association studies. *Plos Genetics*, 5(6): e1000529. doi:10.1371/journal.pgen.1000529.

Howie, B., Marchini, J., & Stephens. (2011). Genotype imputation with thousands of genomes. *G3: Genes, Genomes, Genetics*, 1: 457-469.

Howie, B. *et al*. (2012). Fast and accurate genotype imputation in genome-wide association studies through pre-phasing. *Nature Genetics*, 44(8): 955-960.

Krawczak, M. (2015). Genotype imputation. *In:eLS.John Wiley & Sons, Ltd:Chichester*.

doi:10.1002/ 9780470015902.a0022399.

Li, N., & Stephens, M. (2003). Modeling linkage disequilibrium and identifying recombination hotspots using single-nucleotide polymorphism data. *Genetics*, 165:2213-2233.

Liu Q., *et al*. (2014). Systematic assessment of imputation performance using the 1000 Genomes reference panels. *Briefings in Bioinformatics*, DOI: 10.1093/bib/bbu035.

Marchini, J., *et al*. (2007). A new multipoint method for genome-wide association studies by imputation of genotypes. *Nature Genetics*, 39(7): 906-913.

Marchini, J., & Howie, B. (2010). Genotype imputation for genome-wide association studies. *Nature Reviews Genetics*, 11: 499-511.

Nothnagel, M., *et al*. (2009). A comprehensive evaluation of SNP genotype imputation. *Human Genetics*, 125:163-171.

Roberts, A., *et al*. (2007). Inferring missing genotypes in large SNP panels using fast nearest-neighbor searches over sliding windows. *Bioinformatics*, 23: i401-i407.

Roth, P. (1994). Missing data: A conceptual review for applied psychologists. Personnel Psychology, 47: 537-560. doi:10.1111/j.1744-6570.1994.tb01738.x.

Rubin, D.B. (1976). Inference and missing data. *Biometrika*, 63(3): 581-592.

Servin, B., & Stephens, M. (2007). Imputation-based analysis of association studies: candidate regions and qualitative traits. *Plos Genetics*, 3(7): e114.doi:10.1371/journal.pgen.0030114.

Sheet, P., & Stephens, M. (2006). A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and

haplotypic phase. *The American Journal of Human Genetics*, 78:629-644.

Wang, Y.N., Cai, Z.P., Stothard, P., *et al*. (2012). Fast accurate missing SNP genotype local imputation. *BMC Research Notes*, 5:404.